ST5201x Finals Cheatsheet

Permutation Test
Let $\mathcal{S}_X = \{X_1, \cdots, X_n\}$ and $\mathcal{S}_Y = \{Y_1, \cdots, Y_m\}$ be two samples and $T(\mathcal{S}_X, \mathcal{S}_Y)$ be a test statistic.

1. Pool samples to form joint sample $\mathcal{S}_{XY} = \{X_1, \cdots, X_n, Y_1, \cdots, Y_m\}$.
2. Randomly permute elements in $\mathcal{S}_{XY}$ and split them into two samples $\mathcal{S}_X^*$ and $\mathcal{S}_Y^*$.
3. Treating $\mathcal{S}_X^*$ and $\mathcal{S}_Y^*$ as the original sample, compute the test statistic $T(\mathcal{S}_X^*, \mathcal{S}_Y^*)$.
4. Repeat the above two procedures several times, record the value of test statistic of each time.

Repeating B times, we obtain B values of the test statistic $T^{*(1)}, \cdots, T^*(B)$. Then we compute the p-value of testing $H_0: F_X = F_Y$ (2-sided) as

$$\text{p-value } = 2 \times \frac{1}{B} \sum_{\ell=1}^{B} I\left(\left|T^{*(\ell)}\right| \geq |T(\mathcal{S}_X, \mathcal{S}_Y)|\right)$$

Mann-Whitney U-test
Rank-sum test $\Rightarrow$ $\{R_1, \ldots, R_n, R_{n+1}, \ldots, R_{n+m}\}$. Assign ranks to all observations. Smaller rank, lower value. Use test statistic $T(X_1, \ldots, Y_m) = \sum_{i=1}^n R_i$ i.e. sum of ranks of x-observations. Under null, $\{R_1, \ldots, R_n\}$ is a random subset of $\{1, \ldots, n+m\}$.

Wilcoxon Rank Sum (Signed Rank) Test
Consider paired samples i.e. $(x_1, y_1), \ldots, (x_n, y_n)$, compute rank of absolute value of differences

$$T(X_1, \ldots, Y_n) = \sum_{i=1}^n \operatorname{sign}(x_i - y_i) R_i$$

where $R_i = \operatorname{rank}(|x_i - y_i|)$. p-value $= Pr(T \geq T_{obs})$. Under null, $T = \sum_{i=1}^n i \cdot Z_i$ where $Z \sim Ber(0.5)$

Empirical CDF

$$\widehat{F}_n(x_0) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x_0)$$

MSE and Integrated MSE (MISE):

$$\operatorname{MSE}(x) = \mathbb{E}\left\{(\hat{p}_n(x) - p(x))^2\right\}$$

$$\operatorname{MISE}(\widehat{p}_n(x)) = \mathbb{E}\left(\int (\widehat{p}_n(x) - p(x))^2\right)$$

$$= \int \mathbb{E}\{\widehat{p}_n(x) - p(x)\}$$

Histogram Density Estimator
Consider the partition of $[0,1]$ into M bins

$$B_1 = [0, \frac{1}{M}), B_2 = [\frac{1}{M}, \frac{2}{M}), \cdots, B_M = \left[\frac{M-1}{M}, 1\right].$$

Then for a given point $x \in B_\ell$, the density estimator from the histogram will be

$$\widehat{p}_n(x) = \frac{\text{No. observations in } B_\ell}{n} \times \frac{1}{\text{length of the bin}}$$

$$= \frac{M}{n} \sum_{i=1}^n I(X_i \in B_\ell).$$

Bias-Variance Tradeoff

$$\operatorname{MSE}(\widehat{p}_n(x)) = \operatorname{bias}^2(\widehat{p}_n(x)) + \operatorname{Var}(\widehat{p}_n(x))$$

$$\leq \frac{L^2}{M^2} + M \cdot \frac{p(x^*)}{n} = 2\left(\frac{Lp(x^*)}{n}\right)^{2/3}$$

taking $M_{\text{opt}} = \left(\frac{n \cdot L^2}{p(x^*)}\right)^{1/3}$.

MISE: Suppose that f' is absolutely continuous and that $\int (f'(u))^2 du < \infty$. Then the MISE

$$R\left(\widehat{f}_n, f\right) = \frac{h^2}{12} \int (f'(u))^2 du + \frac{1}{nh} + o\left(h^2\right) + o\left(\frac{1}{n}\right)$$

The value h^* that minimizes the MISE is

$$h^* = \frac{1}{n^{1/3}} \left(\frac{6}{\int (f'(u))^2 du}\right)^{1/3}$$

With this choice of binwidth,

$$R\left(\widehat{f}_n, f\right) \sim \frac{C}{n^{2/3}}, C = (3/4)^{2/3} \left(\int (f'(u))^2 du\right)^{1/3}$$

Kernel Density Estimator
Given a kernel K and bandwidth $h > 0$, the kernel density estimator is defined to be

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

MISE Bound: $R = O(n^{-4/5})$ where $h^* = Cn^{-1/5}$
For smooth densities and Normal kernel, use the bandwidth $h_n = \frac{1.06\widehat{\sigma}}{n^{1/5}}$ where $\widehat{\sigma} = \min\left\{s, \frac{Q}{1.34}\right\}$.

Nadaraya-Watson Kernel Regression
The Nadaraya-Watson kernel estimator is defined by

$$\widehat{r}_n(x) = \sum_{i=1}^n \ell_i(x) Y_i$$

where K is a kernel, and the weights $\ell_i(x)$ are given by

$$\ell_i(x) = \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x - x_j}{h}\right)}.$$

MISE Bound: Risk of the estimator, $R(\widehat{r}_n, r) =$

$$\frac{h_n^4}{4} \left(\int x^2 K(x) dx\right)^2 \int \left(r''(x) + 2r'(x) \frac{f'(x)}{f(x)}\right)^2 dx$$

$$+ \frac{\sigma^2 \int K^2(x) dx}{n h_n} \int \frac{1}{f(x)} dx + o\left(n h_n^{-1}\right) + o\left(h_n^4\right)$$

as $h_n \to 0$ and $nh_n \to \infty$.

Leave-One-Out Cross Validation (LOOCV)
The LOOCV score is defined by

$$\mathrm{CV} = \widehat{R}(h) = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \widehat{r}_{(-i)}(x_i)\right)^2$$

where $\widehat{r}_{(-i)}$ is the estimator obtained by omitting the i-th pair (x_i, Y_i).

Theorem: Let $\widehat{r}_n$ be a linear smoother. Then the LOOCV score $\widehat{R}(h)$ can be written as

$$\widehat{R}(h) = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i - \widehat{r}_n(x_i)}{1 - L_{ii}}\right)^2$$

where $L_{ii} = \ell_i(x_i)$ is the i-th diagonal element of the smoothing matrix L.

Theorem: The cross-validation estimator of risk is

$$\widehat{J}(h) = \int \left(\widehat{f}_n(x)\right)^2 dx - \frac{2}{n} \sum_{i=1}^n \widehat{f}_{(-i)}(X_i)$$

where $\widehat{f}_{(-i)}$ is the density estimator obtained after removing the i-th observation.

Bootstrap Distribution
Given statistic $T(\cdot)$, and data $\{X_1, \ldots, X_n\}$, want to estimate distribution of $T(X_1, \ldots, X_n)$.

- Simulate B bootstrap samples $\{X_1^*, \ldots, X_n^*\}$ (sampling with replacement)
- Compute $T^* = T(X_1^*, \ldots, X_n^*)$ for ea. sample $\Rightarrow \{T_1^*, \ldots, T_B^*\}$

Then the bootstrap estimate of variance is given by

$$\hat{se}_{\text{boot}}^2 = \frac{1}{B} \sum_{b=1}^B \left(T_b^* - \frac{1}{B} \sum_{\ell=1}^B T_\ell^*\right)^2$$

Bootstrap Confidence Intervals

1. Normal Interval
Assume $\mathrm{T_n}$ approximately normal. Then, distribution of $T_n - \mathbb{E}T_n \approx \mathcal{N}(0, \operatorname{Var}_F(T_n)) \approx \mathcal{N}(0, \hat{se}_{\text{boot}}^2)$.

$$\mathrm{CI}_\alpha = (T_n - z_{\alpha/2} \hat{se}_{\text{boot}}, T_n + z_{\alpha/2} \hat{se}_{\text{boot}})$$

Poor coverage if dist. of T_n is not close to normal.

2. Pivotal Interval

$$C_n = \left(2\widehat{\theta}_n - \widehat{\theta}^*_{((1-\alpha/2)B)}, 2\widehat{\theta}_n - \widehat{\theta}^*_{((\alpha/2)B)}\right).$$

Note: Pointwise, asymptotic confidence interval.

3. Studentized Pivotal

$$\left(T_n - z^*_{1-\alpha/2} \widehat{\mathrm{se}}_{\text{boot}}, T_n - z^*_{\alpha/2} \widehat{\mathrm{se}}_{\text{boot}}\right)$$

where z^*_β is the β quantile of $Z^*_{n,1}, \ldots, Z^*_{n,B}$ and

$$Z^*_{n,b} = \frac{T^*_{n,b} - T_n}{\widehat{\mathrm{se}}^*_b}.$$

4. Percentile Interval

$$C_n = \left(T^*_{(B\alpha/2)}, T^*_{(B(1-\alpha/2))}\right),$$

Direct Sampling (Inverse CDF Transform)

$$\Pr(F^{-1}(U) \leq x) = \Pr(U \leq F(x)) = F(x)$$

e.g. Let $U = F(x) = 1 - e^{-x}$ where $U \sim Unif(0,1)$.

$$\Rightarrow X = -ln(1-U) \sim Exp(1)$$

Rejection Sampling
Goal: Sample from unknown density $f(\theta)$. Suppose we have a known function $h(\theta) \propto f(\theta)$.

$$h(\theta) \to \text{target density}$$

$$g(\theta) \to \text{proposal density}$$

Need to find M such that

$$h(\theta) \leq Mg(\theta) \iff M \geq \frac{h(\theta)}{g(\theta)}$$

Optimal $M = \sup_\theta \frac{h(\theta)}{g(\theta)}$ (i.e. maximized over θ)

1. Generate $\theta^{(s)} \sim g(\theta)$ (proposal), and $U \sim Unif(0,1)$
2. If $U < \frac{h(\theta)}{Mg(\theta)}$, accept $\theta^{(s)}$, else reject.

Taking $h(\theta) = cf(\theta)$ for some constant c, then

$$\text{Accept Rate} = \frac{\int h(\theta)}{\int Mg(\theta) d\theta} = \frac{c}{M}$$

Importance Sampling
Let $f(x) = h(x)/c$ where h known, c unknown. Want to est $\mathbf{E}_f(k(x))$ for some function k. Use proposal density g to estimate:

$$\hat{E}_f(k(x)) = \frac{\frac{1}{n} \sum w(x_i) k(x_i)}{\frac{1}{n} \sum w(x_i)} \Rightarrow w(x) = \frac{h(x)}{g(x)}$$

$$ESS = \frac{1}{\sum \hat{w}_i^2} \Rightarrow \hat{w}_i = \frac{w(x_i)}{\sum w(x_i)}$$

where ESS denotes effective sample size.

Markov Chains
Denote the transition probability from state i to state j at time $t+1$ by

$$p_{ij} = P(X_{t+1} = j \mid X_t = i)$$

Given K states, the transition matrix is given by

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{KK} \end{pmatrix}$$

where each row represents a distribution (sums to one). Let the row vector μ_t be the distribution of X_t :

$$\mu_t(x) = \mathbf{P}\{X_t = x\} \quad \text{for all } x \in \Omega.$$

By conditioning on the possible predecessors of the $(t+1)$-st state, we see that

$$\mu_t(y) = \sum_{x \in \Omega} \mathbf{P}\{X_{t-1} = x\} P(x,y)$$

$$= \sum_{x \in \Omega} \mu_{t-1}(x) P(x,y) \quad \text{for all } y \in \Omega.$$

Rewriting this in vector form gives

$$\mu_t = \mu_{t-1} P = \cdots = \mu_0 P^t \quad \text{(matrix power)}$$

Stationary Distribution
Suppose $\mu_t \to \pi$ converges to fixed distribution, then

$$\pi P = \pi \quad \text{(stationary distribution)}$$

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P(x,y) \quad \text{for all } y \in \Omega$$

Detailed Balance Equations:

$$\pi(x) P(x,y) = \pi(y) P(y,x) \quad \text{for all } x, y \in \Omega$$

Note: Detailed balance implies balance.

Convergence

- A Markov chain P is **irreducible** if for all pair of states i, j, there exists a $t > 0$ such that $p_{ij}(t) > 0$.
- A state is periodic if the MC can only return to it at regular intervals. A chain is **aperiodic** if it has no periodic states.

Theorem: If P is irreducible and aperiodic, then μ_t converges to unique stationary distribution.

Simple Random Walk (Graphs)
Define simple random walk on graph, G to be the Markov chain with state space V and transition matrix

$$P(x,y) = \begin{cases} \frac{1}{\deg(x)} & \text{if } y \sim x \\ 0 & \text{otherwise} \end{cases}$$

with stationary distribution given by

$$\pi(y) = \frac{\deg(y)}{2|E|} \quad \text{for all } y \in \Omega$$

where $|E|$ denotes no. of edges.

Markov Chain Monte-Carlo
Suppose we want to modify simple random walk to achieve a specified target distribution. Modify RW with rejection sampling.

New Transition Probabilities:

$$Q(x, y) = \begin{cases} P(x, y)a(x, y) & \text{if } y \neq x \\ 1 - \sum_{z \neq x} P(x, z)a(x, z) & \text{if } y = x \end{cases}$$

Acceptance Probability:

$$a(x, y) = \min\left\{\frac{\pi(y)P(y, x)}{\pi(x)P(x, y)}, 1\right\}$$

Metropolis-Hastings Algorithm

1. Choose initial $\theta^{(0)}$, for which $p\left(\theta^{(0)} \mid y\right) > 0$.
2. For $t = 1, 2, \ldots$:
   - Sample a proposal θ^* from the proposal distribution $g\left(\theta^* \mid \theta^{(t-1)}\right)$
   - Compute the ratio

$$r = \frac{f\left(\theta^*\right) g\left(\theta^{(t-1)} \mid \theta^*\right)}{f\left(\theta^{(t-1)}\right) g\left(\theta^* \mid \theta^{(t-1)}\right)}$$

$$\text{Set } \theta^{(t)} = \begin{cases} \theta^* & \text{with prob. } \min\{r, 1\} \\ \theta^{(t-1)} & \text{otherwise.} \end{cases}$$

where f denotes stationary distribution

Autocorrelation of stationary Markov Chain

$$\rho_L = \frac{\sum_{t=0}^{T-L}(f(\theta_t) - \overline{f(\theta)})(f(\theta_{t+L}) - \overline{f(\theta)})}{\sum_{t=0}^{T}(f(\theta_t) - \overline{f(\theta)})^2}$$

Effective Sample Size (ESS)
Let T be number of samples and unknown but stationary $f(\theta_t)$ be drawn from some MCMC algorithm. ESS is

$$T \times (1 + 2\sum_{L}^{\infty} \rho_L)^{-1}$$

Bernoulli Distribution, $X \sim Ber(p)$

$$P(X = x) = \begin{cases} p^x(1-p)^x & \text{if } x = 0 \text{ or } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

Its MGF, $M(t) = 1 - p + pe^t$.

Binomial Distribution, $Y \sim Bin(n, p)$

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

MGF: $(1 - p + pe^t)^n$

Poisson Distribution

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$E[X] = Var(X) = \lambda$, and MGF $M(t) = e^{\lambda(e^t - 1)}$.

Geometric Distribution, $X \sim Geom(p)$
Number of trials until the first success.

$$P(X = k) = (1-p)^{k-1} p$$

$E[X] = 1/p$ and $Var(X) = (1-p)/p^2$, and MGF, $M(t) = pe^t/(1-(1-p)e^t)$.

Note: (Tail Probability Formula)

$$P(X \geq K) = \sum_{i=k}^{\infty} p(1-p)^{i-1} = p\frac{(1-p)^{k-1}}{1-(1-p)} = (1-p)^{k-1}$$

Negative Binomial, $X \sim NegBin(r, p)$
Suppose that a sequence of independent trials performed such that there are r successes in total.

$$P(X = k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r$$

Equivalently, $X = Y_1 + \cdots + Y_r$ where $Y_i's$ are i.i.d. $Geom(p)$. $E[X] = r/p$ and $Var(x) = r(1-p)/p^2$

Hypergeometric Distribution
Given population of size N, of which M are of type I, $N - M$ are of type II, want to draw n samples without replacement, then

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \qquad 0 \leq k \leq n$$

where X is the number of Type I selected.

Uniform Random Variable, $X \sim Unif(a, b)$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Note that for any $(x, y) \in [a, b]$:

$$P(x < X < y) = \frac{y - x}{b - a}$$

Mean, Variance and MGF:

$$\text{E}[X] = \frac{a+b}{2}, \text{Var}(X) = \frac{(b-a)^2}{12}, M(t) = \frac{e^t - 1}{t}$$

CDF:

$$F_X(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{if } x > b \end{cases}$$

Exponential Random Variable
$X \sim Exp(\lambda)$ where $\lambda > 0$ if it has the pdf

$$f_X(x) = \lambda e^{-\lambda x} \quad x \geq 0 \quad (0 \text{ otherwise})$$

whereby its mean and variance are given by $1/\lambda$ and $1/\lambda^2$ respectively. MGF is $\lambda/(\lambda - t)$.

Method of Moments Estimators
The k-th moment of a probability law is defined as

$$\mu_k = \text{E}[X^k]$$

If $X_1, X_2, \ldots, X_n$ are iid random variables from that distribution, the k-th sample moment is defined as

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

where $\hat{\mu}_k$ can be taken as an estimate of μ_k.

MLE for Multinomial Cell Probabilities
The likelihood is the joint frequency function

$$\text{lik}\,(p_1, \ldots, p_m) = \frac{n!}{\prod_{i=1}^{m} x_i!}\prod_{i=1}^{m} p_i^{x_i}$$

Note: Marginal distribution of each X_i is $Bin\,(n, p_i)$.

The log likelihood is

$$l\,(p_1, \ldots, p_m) = \log n! - \sum_{i=1}^{m}\log x_i! + \sum_{i=1}^{m} x_i \log p_i$$

Bootstrap Procedure - Multinomial Example
Boostrap procedure for approximating the sampling distributions of $\hat{\theta}$:

1. Assume multinomial distribution with $\hat{\theta}$ provides a good fit to the data
2. Simulate N random samples from multinomial distribution with corresponding probabilities p_1, p_2, p_3 and $n = 1029$.
3. For ea. random sample, calculate MLE, θ^* of θ
4. Use the N values of θ^* to approximate the sampling distributions of $\hat{\theta}$

The standard error of $\hat{\theta}$ can be estimated using

$$s_{\hat{\theta}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta_i^* - \bar{\theta})^2} \quad \text{where} \quad \bar{\theta} = \frac{1}{N}\sum_{i=1}^{N}\theta_i^*$$

Consistency
Let $\hat{\theta}_n$ be an estimate of a parameter θ_0 based on a sample of size n. $\hat{\theta}_n$ is said to be consistent in probability if $\hat{\theta}_n$ converges in probability to θ_0 as n approaches infinity. That is, for $\epsilon > 0$,

$$P(|\hat{\theta}_n - \theta_0| > \epsilon) \to 0 \text{ as } n \to \infty$$

Fisher Information (Lemma A)

$$I(\theta) = \text{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2$$

$$= -\,\text{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$$

Large Sample Theory for MLE
Let $\hat{\theta}$ denote the MLE of θ_0. The probability distribution of

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$$

tends to a standard normal distribution. Therefore, the asymptotic variance of the MLE is

$$\frac{1}{nI(\theta)} = -\frac{1}{\text{E}[l''(\theta_0)]}$$

Efficiency
The efficiency of two estimators, $\hat{\theta}_0, \hat{\theta}_1$ is given as

$$\text{eff}(\hat{\theta}_0, \hat{\theta}_1) := \text{Var}(\hat{\theta}_1)/\,\text{Var}(\hat{\theta}_0)$$

If the effficiency is smaller than 1, then $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_0)$

Cramer-Rao Lower Bound
Under smoothness assumptions of a $f(x|\theta)$ for a statistic $T := t(X_1, \cdots, X_n)$

$$\text{Var}(T) \geq \frac{1}{nI(\theta)}$$

This gives the lower bound for the variance of any estimator of θ.

Sufficiency
A statistic $T(X_1, \cdots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, \cdots, X_n$ given $T = t$ does not depend on θ for any value of t. If T is sufficient for θ, the MLE for θ is a function only of T.

Factorization Theorem
The statistic $T(X_1, \cdots, X_n)$ is sufficient for a parameter θ iff the joint pdf factorises in the form

$$f(\vec{x}|\theta) = g(T(\vec{x}), \theta)h(\vec{X})$$

Generalized Likelihood Ratio Test
Denote the null and alternative hypotheses as $H_0 : \theta \in \omega_0$ and $H_1 : \theta \in \omega_1$ respectively, where ω_0, ω_1 are disjoint and subsets of Ω, the sample space. The generalised likelihood ratio test statistic is

$$\Lambda^* = \frac{\max_{\theta\in\omega_0} L(\theta)}{\max_{\theta\in\omega_1} L(\theta)}$$

For simplicity, we define Λ such that $\Lambda = \min(\Lambda^*, 1)$.

$$\Lambda = \frac{\max_{\theta\in\omega_0} L(\theta)}{\max_{\theta\in\Omega} L(\theta)}$$

Then, the generalised likelihood test rejects for $\Lambda \leq \lambda_0$, where $P(\Lambda \leq \lambda_0|H_0) = \alpha$.

Distribution of $-2\log\Lambda$
As the sample size $n \to \infty$, null distribution of $-2\log\Lambda$ tends to a chi-square distribution with degrees of freedom df as

$$df = \dim\Omega - \dim\omega_0$$

Rejecting for small Λ is then also rejecting for large $-2\log\Lambda$.

Likelihood Ratio Test (LRT)
In the case of the simple alternative hypothesis, simply define Λ directly.

$$\Lambda = \frac{L(\theta|H_0)}{L(\theta|H_1)}$$

Pearson Chi-square Test
The Pearson chi-square test is asymptotically equal to the GLRT. The test statistic for a multinomial distributed r.v. is

$$X^2 = \sum_{i=1}^{m}\frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{m}\frac{(x_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})}$$

Where $X^2 \sim \chi^2_{m-k-1}$, k is the number of values of the multinomial distribution.

Bayesian Inference
Let unknown parameter Θ be a random variable with prior distribution. The posterior distribution is given by:

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X,\Theta}(x, \theta)}{f_X(x)}$$

$$= \frac{f_{X|\Theta}(x \mid \theta)f_\Theta(\theta)}{\int f_{X|\Theta}(x \mid \theta)f_\Theta(\theta)d\theta}$$

i.e. Posterior density $\propto$ Likelihood $\times$ Prior density

Useful Results: Beta Distribution

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}, \quad 0 \leq x \leq 1$$

$$E(X) = \frac{a}{a+b} \quad , \quad Var(X) = \frac{ab}{(a+b)^2(a+b+)}$$

Beta Integral

$$\int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

where

$$\Gamma(n) = (n-1))! \quad , \quad \Gamma(n+1) = n\Gamma(n)$$